CLAIMS

1. Apparatus for processing image data and sound data, comprising:

an image processor for processing image data recorded by at least one camera showing the movements of a plurality of people to track each person in three dimensions;

a sound processor for processing sound data to determine the direction of arrival of the sound;

a speaker identifier for determining which of the people is speaking based on the result of the processing performed by the image processor and the result of the processing performed by the sound processor; and

a voice recognition processor for processing the received sound data to generate text data therefrom in dependence upon the result of the processing performed by the speaker identifier.

2. Apparatus according to claim 1, wherein the voice recognition processor includes a store for storing respective voice recognition parameters for each of the people, and a selection processor for selecting the voice recognition parameters to be used to process the sound data in dependence upon the person determined to be

speaking by the speaker identifier.

3. Apparatus according to claim 1, wherein the image processor is arranged to track each person by processing the image data using camera calibration data defining the position and orientation of each camera from which image data is processed.

4. Apparatus according to claim 1, wherein the image processor is arranged to track each person by tracking each person's head.

5. Apparatus according to claim 1, wherein the image processor is arranged to process the image data to determine where at least each person who is speaking is looking.

6. Apparatus according to claim 1, wherein the speaker identifier is arranged to identify a person who is speaking in a given frame of the received image data using the results of the processing performed by the image processor and the sound processor for at least one other frame if the speaker cannot be identified using the results of the processing performed by the image processor and the sound processor for the given frame.

7. Apparatus according to claim 1, further comprising a database for storing at least some of the received image data, the sound data, the text data produced by the voice recognition processor and viewing data defining where at least each person who is speaking is looking, the database being arranged to store the data such that corresponding text data and viewing data are associated with each other and with the corresponding image data and sound data.

8. Apparatus according to claim 7, further comprising a data compressor for compressing the image data and the sound data for storage in the database.

9. Apparatus according to claim 8, wherein the data compressor comprises a data encoder for encoding the image data and the sound data as MPEG data.

10. Apparatus according to claim 7, further comprising a gaze data generator for generating data defining, for a predetermined period, the proportion of time spent by a given person looking at each of the other people during the predetermined period, and wherein the database is arranged to store the data so that it is associated with the corresponding image data, sound data, text data and

viewing data.

11. Apparatus according to claim 10, wherein the predetermined period comprises a period during which the given person was talking.

12. Apparatus for processing image data and sound data, comprising:

an image processor for processing image data recorded by at least one camera showing the movements of a plurality of people to track each person in three dimensions;

a sound processor for processing sound data to determine the direction of arrival of the sound; and

a speaker identifier for determining which of the people is speaking based on the result of the processing performed by the image processor and the result of the processing performed by the sound processor.

13. Apparatus according to claim 12, wherein the image processor is arranged to track each person by processing the image data using camera calibration data defining the position and orientation of each camera from which image data is processed.

14. Apparatus according to claim 12, wherein the image processor is arranged to track each person by tracking each person's head.

15. Apparatus according to claim 12, wherein the image processor is arranged to process the image data to determine where at least each person who is speaking is looking.

16. Apparatus according to claim 12, wherein the speaker identifier is arranged to identify a person who is speaking in a given frame of the received image data using the results of the processing performed by the image processor and the sound processor for at least one other frame if the speaker cannot be identified using the results of the processing performed by the image processor and the sound processor for the given frame.

17. A method of processing image data and sound data, comprising:

an image processing step comprising processing image data recorded by at least one camera showing the movements of a plurality of people to track each person in three dimensions;

a sound processing step comprising processing sound

data to determine the direction of arrival of the sound;

a speaker identification step comprising determining which of the people is speaking based on the result of the processing performed in the image processing step and the result of the processing performed in the sound processing step; and

a voice recognition processing step comprising processing the received sound data to generate text data therefrom in dependence upon the result of the processing performed in the speaker identification step.

18. A method according to claim 17, wherein, the voice recognition processing step includes selecting, from stored respective voice recognition parameters for each of the people, the voice recognition parameters to be used to process the sound data in dependence upon the person determined to be speaking in the speaker identification step.

19. A method according to claim 17, wherein, in the image processing step, each person is tracked by processing the image data using camera calibration data defining the position and orientation of each camera from which image data is processed.

20. A method according to claim 17, wherein, in the image processing step, each person is tracked by tracking the person's head.

21. A method according to claim 17, wherein, in the image processing step, the image data is processed to determine where at least each person who is speaking is looking.

22. A method according to claim 17, wherein, in the speaker identification step, a person who is speaking in a given frame of the received image data is identified using the results of the processing performed in the image processing step and the sound processing step for at least one other frame if the speaker cannot be identified using the results of the processing performed in the image processing step and the sound processing step for the given frame.

23. A method according to claim 17, further comprising the step of generating a signal conveying the text data generated in the voice recognition processing step.

24. A method according to claim 17, further comprising the step of storing in a database at least some of the

received image data, the sound data, the text data produced in the voice recognition processing step and viewing data defining where at least each person who is speaking is looking, the data being stored in the database such that corresponding text data and viewing data are associated with each other and with the corresponding image data and sound data.

25. A method according to claim 24, wherein the image data and the sound data are stored in the database in compressed form.

26. A method according to claim 25, wherein the image data and the sound data are stored as MPEG data.

27. A method according to claim 24, further comprising the steps of generating data defining, for a predetermined period, the proportion of time spent by a given person looking at each of the other people during the predetermined period, and storing the data in the database so that it is associated with the corresponding image data, sound data, text data and viewing data.

28. A method according to claim 27, wherein the predetermined period comprises a period during which the

given person was talking.

29. A method according to claim 24, further comprising the step of generating a signal conveying the database with data therein.

30. A method according to claim 29, further comprising the step of recording the signal either directly or indirectly to generate a recording thereof.

31. A method of processing image data and sound data, comprising:

an image processing step comprising processing image data recorded by at least one camera showing the movements of a plurality of people to track each person in three dimensions;

a sound processing step comprising processing sound data to determine the direction of arrival of the sound; and

a speaker identification step comprising determining which of the people is speaking based on the result of the processing performed in the image processing step and the result of the processing performed in the sound processing step.

32. A method according to claim 31, wherein, in the image processing step, each person is tracked by processing the image data using camera calibration data defining the position and orientation of each camera from which image data is processed.

33. A method according to claim 31, wherein, in the image processing step, each person is tracked by tracking the person's head.

34. A method according to claim 31, wherein, in the image processing step, the image data is processed to determine where at least each person who is speaking is looking.

35. A method according to claim 31, wherein, in the speaker identification step, a person who is speaking in a given frame of the received image data is identified using the results of the processing performed in the image processing step and the sound processing step for at least one other frame if the speaker cannot be identified using the results of the processing performed in the image processing step and the sound processing step for the given frame.

36. A method according to claim 31, further comprising the step of generating a signal conveying the identity of the speaker identified in the speaker identification step.

37. A storage device storing instructions for causing a programmable processing apparatus to become configured as an apparatus as set out in at least one of claims 1 and 12.

38. A storage device storing instructions for causing a programmable processing apparatus to become operable to perform a method as set out in at least one of claims 17 and 31.

39. A signal conveying instructions for causing a programmable processing apparatus to become configured as an apparatus as set out in at least one of claims 1 and 12.

40. A signal conveying instructions for causing a programmable processing apparatus to become operable to perform a method as set out in at least one of claims 17 and 31.

41. Apparatus for processing image data and sound data, comprising:

image processing means for processing image data recorded by at least one camera showing the movements of a plurality of people to track each person in three dimensions;

sound processing means for processing sound data to determine the direction of arrival of the sound;

speaker identification means for determining which of the people is speaking based on the result of the processing performed by the image processing means and the result of the processing performed by the sound processing means; and

voice recognition processing means for processing the received sound data to generate text data therefrom in dependence upon the result of the processing performed by the speaker identification means.

42. Apparatus for processing image data and sound data, comprising:

image processing means for processing image data recorded by at least one camera showing the movements of a plurality of people to track each person in three dimensions;

sound processing means for processing sound data to

determine the direction of arrival of the sound; and

speaker identification means for determining which of the people is speaking based on the result of the processing performed by the image processing means and the result of the processing performed by the sound processing means.

43. Apparatus for processing image data and sound data, comprising:

an image processor for processing image data recorded by at least one camera showing the movements of a plurality of people to determine where each person is looking and to determine which of the people is speaking based on where the people are looking; and

a sound processor for processing sound data defining words spoken by the people to generate text data therefrom in dependence upon the result of the processing performed by the image processor.

44. Apparatus according to claim 43, wherein the sound processor includes a store for storing respective voice recognition parameters for each of the people, and a selection processor for selecting the voice recognition parameters to be used to process the sound data in dependence upon the person determined to be speaking by

the image processor.

45. Apparatus according to claim 43, wherein the image processor is arranged to determine where each person is looking by processing the image data using camera calibration data defining the position and orientation of each camera from which image data is processed.

46. Apparatus according to claim 43, wherein the image processor is arranged to determine where each person is looking by processing the image data to track the position and orientation of each person's head in three dimensions.

47. Apparatus according to claim 43, wherein the image processor is arranged to determine which person is speaking based on the number of people looking at each person.

48. Apparatus according to claim 47, wherein the image processor is arranged to generate a value for each person defining at whom the person is looking and to process the values to determine the person who is speaking.

49. Apparatus according to claim 43, wherein the image

processor is arranged to determine that the person who is speaking is the person at whom the most other people are looking.

50. Apparatus according to claim 43, further comprising a database for storing the image data, the sound data, the text data produced by the sound processor and viewing data defining where each person is looking, the database being arranged to store the data such that corresponding text data and viewing data are associated with each other and with the corresponding image data and sound data.

51. Apparatus according to claim 50, further comprising a data compressor for compressing the image data and the sound data for storage in the database.

52. Apparatus according to claim 51, wherein the data compressor comprises a data encoder for encoding the image data and the sound data as MPEG data.

53. Apparatus according to claim 50, further comprising a gaze data generator for generating data defining, for a predetermined period, the proportion of time spent by a given person looking at each of the other people during the predetermined period, and wherein the database is

arranged to store the data so that it is associated with the corresponding image data, sound data, text data and viewing data.

54. Apparatus according to claim 53, wherein the predetermined period comprises a period during which the given person was talking.

55. Apparatus for processing image data, comprising:

a receiver for receiving image data recorded by at least one camera showing the movements of a plurality of people; and

an image processor for processing the image data to determine where each person is looking and to determine which of the people is speaking based on where the people are looking.

56. Apparatus according to claim 55, wherein the image processor is arranged to determine where each person is looking by processing the image data using camera calibration data defining the position and orientation of each camera from which image data is processed.

57. Apparatus according to claim 55, wherein the image processor is arranged to determine where each person is

looking by processing the image data to track the position and orientation of each person's head in three dimensions.

58. Apparatus according to claim 55, wherein the image processor is arranged to determine which person is speaking based on the number of people looking at each person.

59. Apparatus according to claim 58, wherein the image processor is arranged to generate a value for each person defining at whom the person is looking and to process the values to determine the person who is speaking.

60. Apparatus according to claim 55, wherein the image processor is arranged to determine that the person who is speaking is the person at whom the most other people are looking.

~~61.~~ A method of processing image data and sound data, comprising:

an image processing step comprising processing image data recorded by at least one camera showing the movements of a plurality of people to determine where each person is looking and to determine which of the

people is speaking based on where the people are looking; and

a sound processing step comprising processing sound data defining words spoken by the people to generate text data therefrom in dependence upon the result of the processing performed in the image processing step.

62. A method according to claim 61, wherein the sound processing step includes selecting, from stored respective voice recognition parameters for each of the people, the voice recognition parameters to be used to process the sound data in dependence upon the person determined to be speaking in the image processing step.

63. A method according to claim 61, wherein, in the image processing step, it is determined where each person is looking by processing the image data using camera calibration data defining the position and orientation of each camera from which image data is processed.

64. A method according to claim 61, wherein, in the image processing step, it is determined where each person is looking by processing the image data to track the position and orientation of each person's head in three dimensions.

65. A method according to claim 61, wherein, in the image processing step, it is determined which person is speaking based on the number of people looking at each person.

66. A method according to claim 65, wherein, in the image processing step, a value is generated for each person defining at whom the person is looking and the values are processed to determine the person who is speaking.

67. A method according to claim 61, wherein, in the image processing step, it is determined that the person who is speaking is the person at whom the most other people are looking.

68. A method according to claim 61, further comprising the step of storing the image data, the sound data, the text data produced in the sound processing step and viewing data defining where each person is looking in a database, the database being arranged to store the data such that corresponding text data and viewing data are associated with each other and with the corresponding image data and sound data.

69. A method according to claim 68, wherein the image data and the sound data are stored in the database in compressed form.

70. A method according to claim 69, wherein the image data and the sound data are stored as MPEG data.

71. A method according to claim 68, further comprising the steps of generating data defining, for a predetermined period, the proportion of time spent by a given person looking at each of the other people during the predetermined period, and storing the data in the database so that it is associated with the corresponding image data, sound data, text data and viewing data.

72. A method according to claim 71, wherein the predetermined period comprises a period during which the given person was talking.

73. A method according to claim 68, further comprising the step of generating a signal conveying the database with data therein.

74. A method according to claim 73, further comprising the step of recording the signal either directly or

indirectly to generate a recording thereof.

75. A method of processing image data, comprising:

receiving image data recorded by at least one camera showing the movements of a plurality of people; and

processing the image data to determine where each person is looking and to determine which of the people is speaking based on where the people are looking.

76. A method according to claim 75, wherein it is determined where each person is looking by processing the image data using camera calibration data defining the position and orientation of each camera from which image data is processed.

77. A method according to claim 75, wherein it is determined where each person is looking by processing the image data to track the position and orientation of each person's head in three dimensions.

78. A method according to claim 75, wherein it is determined which person is speaking based on the number of people looking at each person.

79. A method according to claim 78, wherein a value is

generated for each person defining at whom the person is looking and the values are processed to determine the person who is speaking.

80. A method according to claim 75, wherein it is determined that the person who is speaking is the person at whom the most other people are looking.

81. A storage device storing instructions for causing a programmable processing apparatus to become configured as an apparatus as set out in at least one of claims 43 and 55.

82. A storage device storing instructions for causing a programmable processing apparatus to become operable to perform a method as set out in at least one of claims 61 and 75.

83. A signal conveying instructions for causing a programmable processing apparatus to become configured as an apparatus as set out in at least one of claims 43 and 55.

84. A signal conveying instructions for causing a programmable processing apparatus to become operable to

perform a method as set out in at least one of claims 61 and 75.

85. Apparatus for processing image data and sound data, comprising:

image processing means for processing image data recorded by at least one camera showing the movements of a plurality of people to determine where each person is looking and to determine which of the people is speaking based on where the people are looking; and

sound processing means for processing sound data defining words spoken by the people to generate text data therefrom in dependence upon the result of the processing performed by the image processing means.

86. Apparatus for processing image data, comprising:

receiving means for receiving image data recorded by at least one camera showing the movements of a plurality of people; and

means for processing the image data to determine where each person is looking and to determine which of the people is speaking based on where the people are looking.